DOCUMENT RESUME

ED 309 128 SO 020 112

AUTHOR Dolbear, F. Trenery, Jr.
TITLE Software for Economics Statistics Courses.
PUB DATE 12 Mar 88
NOTE 39p.; Paper presented at the Annual Convention of the Eastern Economic Association (14th, Boston, MA, March 10-12, 1988).
PUB TYPE Book/Product Reviews (072) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Instruction; *Computer Software; *Economics; *Economics Education; Higher Education; *Instructional Improvement; Instructional Innovation; Microcomputers; Regression (Statistics); *Statistics
IDENTIFIERS Apple Macintosh; Stat Helper

ABSTRACT

Three aspects to be considered when teaching a one-semester beginning economics statistics course are coverage, mastery, and applications. There is a difference between coverage and mastery. Moreover, mastery is not an end in itself; instructors must consider how statistics courses will influence students' approaches to other subjects and applications. In principle, computer activities can be designed and implemented to improve any and all of these three goals. The HyperCard software for the Macintosh computer should result in an important advance in the interface between computer and user. This will be valuable for tutorial programs. Cognitive scientists are designing software which analyzes student solutions to standard problems by inferring a student's intentions from the details of her/his solution and then offering diagnostic assistance. Programs like "Stat Helper" (briefly discussed) for the Macintosh allow students to interact with the computer in solving a variety of problems. Students can learn about regressions better through hands-on experience on personal or mainframe computers. Computer experiments can exhibit a variety of points about regression applications. Computers can expand coverage and make applications more accessible to the average student. Students must develop some sense about what questions regressions can and cannot be expected to answer. Four examples, including graphs and statistical data, are given: automobile weight and fuel mileage, polynomial (quadratic), omitted independent variable, and logarithmic relationship. Nine references and numerous tables and graphs are provided. (GEA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED309128

Software for Economics Statistics Courses

F. Trenery Dolbear, Jr.
Department of Economics
Brandeis University
Waltham, MA 02254
(617) 736-2254

Presented at Eastern Economic Association's 14th Annual Convention, Boston, MA, March 10-12, 1988.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

F. TRENERY DOLBEAR JR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

1. Introduction

My instructions for this session—I should admit that these instructions were "self-written," so as to spare our organizer blame should someone wish to take exception to the views expressed below—call for some information on and critique of software currently available for use in Economics statistics courses. I interpret my mission not only to tell you about some of the products which have come across my desk[1] but also to

---

[1]I say "come across my desk" intentionally for I have not conducted an exhaustive survey of what's out there.

---

offer my opinion more generally on how we ought to use computers in our quest for more effective economics statistics courses. This will be a follow-up on some ideas and product reviews I offered about a year ago in a paper for the Social Science Microcomputer Review [4]. The views I state here are consistent with the outlook I held a year ago, but except where repetition is essential for the argument, I'll try to avoid a restatement of views already in print.

To begin at the beginning I think it is first necessary to ask

what we are trying to accomplish in a one-semester beginning economics statistics course.[2] I will discuss three aspects of the

---

[2] The second course, which in most incarnations is probably titled "Econometrics," is not a major focus of this paper. As far as I have been able to make out, the primary use of computers in that course involves a regression package. My colleagues, and probably yours as well, all have their favorites and they aren't looking for advice. Of course some of the promising pedagogical ideas involving computers could be applied in the second course as well as the first. Indeed in a Journal of Economic Education paper [7] Lawrence Pulley and I outlined some computer exercises we developed for exhibiting some of the problems which result where one or another of the standard textbook least squares regression model assumptions are not met, but the analyst uses least squares anyway. I have never managed to get far enough in the first statistics course to use these exercises, but a colleague has told me that such activities make a nice addition to the early weeks of the Econometrics course.

---

answer to that question: coverage, mastery, and applications.

I answer the *coverage* part of this question by seeking a common denominator from the statistics textbooks that line my shelves. They all seem to have a little descriptive statistics, some probability topics, introductions to the binomial and normal distributions, the distribution of the sample mean (including some discussion of the central limit theorem) and perhaps of differences in two means, point and interval estimation, hypothesis testing, and regression, both two variable and more-than-two (but often no more than three). Occasionally some analysis of variance is also included; this is useful in understanding R^2s and F tests which tend to show up in the more advanced regression chapters. There are other, more advanced topics as well but my hunch, buttressed with whatever anecdotal evidence I have been able to muster, is that most of those topics are omitted in the beginning course.

My view that textbooks provide a reliable guide to what teachers teach has generally been supported by casual interviews. Whenever I have the opportunity, I quiz instructors and students from other institutions about the content of their courses. The answers are mostly in agreement with the above list. Of course not all of us are at all times successful in accomplishing what we

set out to do. Actual coverage isn't always identical to what might be anticipated by taking the course description at its face value.[3]

---

[3] For example many undergraduates who go off hither and yon to take statistics in summer school return in the fall having had only a very limited exposure to regression—the catalogue's promises for the coverage of that topic notwithstanding. I don't have a lot of wisdom on how to insure that our ex post realizations live up to our ex ante promises, but I can recommend trying to live by a daily schedule set down at the beginning of the course and then amended if necessary in the light of experience. For economists, regression is not the best topic to squeeze in the traditonal end-of-semester compression of the course time-table. In my opinion the best way to avoid that squeeze is to anticipate it by being constantly aware of the schedule so that one can choose in advance what topics ought to be rushed. This is clearly more rational than placing the crunch on regression, by default, simply because it happens to come at the end of the semester.

Of course coverage need not be restricted to what is in the textbooks. We should consider the possibility, or perhaps likelihood, that the general availability of high powered, user friendly, personal[4] computers will make the teaching of some

---

[4]Although I, and I believe many students, have found that the "personal" (in the sense of at your own desk, in your everyday working environment) aspect of PC's has markedly improved learning, substantial improvements over the older mainframe, shared computing system, can be had even from PC's located in a university laboratory because of the user-friendly operating system, improved graphics, and accessible printing resources.

---

important topics feasible that are difficult if not impossible to deliver within the customary textbook-lecture-problem set format of instruction. For some thoughts on this point I refer you to earlier papers ([4], [7]) which discuss some general principles and provide examples. The bottom line is that as PC's become widely-available, *coverage* should be changed and for the better.

Of course there is a difference between "covering" a topic

and *mastering* that topic as all of us who read final examination answers know. This suggests we ought to reexamine the effectiveness of our teaching strategy (textbook, lectures, problem sets, quizzes, computer exercises, etc.) as well as our selection of topics.[5]

---

[5]I don't wish to underestimate the importance of the learning attitudes students bring to our courses by focusing on "teaching." Such attitudes are at least as important, but they are likely to be a variable that instructors cannot easily control.

Moreover mastery of textbook materials isn't an end in itself; we need to have in mind how our courses in statistics will influence our students' approaches to other subjects or more generally their 'world view.' In an earlier paper [4] I suggested that other courses ought to take a hand in teaching the *applications* of concepts first formally taught in statistics. Too often we wave our hands when we lecture on statistical methods and results in journal articles or advanced textbooks. And in those few cases where our intentions vis-a-vis statistical arguments are on target we allow ourselves to be dissuaded by

student protestations that they cannot be expected to understand such material—they have forgotten whatever statistics they once knew (and they don't admit that they ever knew very much). I think most students are excessively modest or else openly disingenuous on this point; beneath the surface, I am confident, lies the necessary infrastructure to comprehend the significance of a regression result or a formal statistical test of a proposition, if only the instructor will work at brushing away the dust and cobwebs.

In principle computer activities can be designed and implemented to improve any and all of the above goals—coverage, mastery, and applications; and in what follows I will discuss strategies and software for improving performance in all three areas.

2. Mastery.

In [4] I offer a list of potential topics normally covered in a semester of working through the conventional statistics course for which I think the computer has a comparative (and I hope even an absolute) advantage. In addition I discussed commercially available software ([9], [3], and [5]) which can provide some of

this *mastery* instruction. I will not repeat the list here. Rather I will descibe three additional approaches which have come to my attention within the last year.

First, the availability of the new software for the Macintosh called HyperCard should result in an important advance in the interface between computer and user. This will be valuable for tutorial programs. For example I have heard of a proposal to place a formula such as the familiar binomial distribution:

$$p(s) = \frac{n!}{s!\,(n-s)!}\,\pi^{s}(1-\pi)^{n-s},\quad s = 0,\ldots,n$$

on the PC screeen. A student, unsure of the significance of one or another of the ingredients of the binomial distribution formula, could request help by clicking (with the mouse-directed cursor) on any part of the formula he didn't understand. This would bring to the screen an explanation or a demonstration of the fine (or perhaps coarse) points surrounding that part of the formula. Then he could continue peeling off layers increasing the details of his knowledge or he could return to the original formula and click on some other part which would trigger a parallel explanation path.

The important pedagogical point is that students would be

able to call up the explanations with whatever detail seems appropriate given their level of comprehension and whenever they feel ready for it. Speed through the exercise and repetition of any parts would be entirely under the user's control. Potentially this is a big improvement over the conventional lecture method. During a lecture students who look out the window at the wrong moment or who are puzzled by a particular part of the instructor's exposition have little hope of staying up to (the instructor-set) speed. With the PC pausing to look out the window from time to time is no problem. Also a student can pick his own study strategy; of course the desirability of student-controlled pedagogy presumes that the student is the best judge of how he can best master the relevant concepts. After I debrief students on dismal examination performances, I find I have some reservations about the wisdom of their study strategy.

Second, cognitive scientists are beginning to design computer programs which can analyze student solutions to standard problems by inferring a student's intentions from the details of his solution and then offering diagnostic assistance. See, for example, Sebrechts and Schooler [8] who describe a program called GIDE (for generalized intention-based diagnostic system)

which they use to analyze errors in a student solution to a paired sample (before and after treatment) t test. The student is given some paired data reporting maintenance intervals for four jet engines before and after the introduction of a new engine oil, and then asked to determine whether or not the treatment is effective. The program is able to "understand" the details of the student's answer including such subtleties as detecting an implicit but logically correct missing step. Also the program can follow an incorrect numerical answer forward to determine that, although the final answer was incorrect (because, for example, early on the student had used variance where standard deviation was called for), the student correctly completed the problem given the early error. This is a promising development and, compared to what we have seen in computer software so far, valuable because of its scarcity.

I do, however, have some reservations about the usefulness of computer applications such as the engine oil t test execution evaluation described above because in my experience the serious problems aren't caused by students making execution mistakes such as using a variance instead of a standard deviation. (It is typically the students who don't put in the study hours who are

confused about execution procedures.) More likely the student doesn't really understand why it should be a paired test to begin with. Or if a student decides to use a paired sample t test, she won't be able to tell you what she must be assuming about the underlying distributions to make her choice appropriate. What we need is some "help" procedures for students who, when confronted with a statistics problem, are likely to start off on the wrong foot. How can they be taught to recognize what model is most appropriate for the problem at hand?

The third, and last, mastery product I will discuss is some software titled "Stat Helper" for the Macintosh; it was developed at Drexel University by Dennis Dunn and Mary Ann Atkinson [6]. This software takes advantage of Macintosh features and allows students to interact with the computer in solving a variety of the traditional first course problems. A decision tree framework allows the student to advance systematically in working through problems and to begin again if necessary.

In general I value a product students can use in a variety of ways; in my experience they usually learn something with such hands on activity. But this learning requires user friendly and generally bug-free programs. I encountered several problems with

this product, however, which might confuse students; I will give three examples. I should state that I have only sampled this software, and I may not be the best judge of its usefulness as a pedagogical aid.

First, there is a little calculation program for the "proportion distribution"—a binomial random variable divided by n, the number of trials. The program represents the proportion as a continuous distribution which will confuse any student who was taught that it is discrete. Incorrect information ought not be dispensed in a computer program—as well as in textbooks and in the classroom.

Second, in performing a statistical test I had to specify a confidence level; I was confronted with a box containing $(1 - \alpha)$. I backspaced once, removing the α, and typed in the usual .05. The program gave me an error message: "Confidence level must be between 0 and 1." I was a little surprised, but I tried removing the $(1 - \alpha)$ and entering .95. That produced the 5% confidence level I was looking for—as well as a strong reaction that this program could hardly be called "user-friendly."

Third, I came upon a "regression puzzle" in an advanced part of the software. The idea is: Type in whatever information

you have about a regression, for example an estimate of a slope coefficient and its standard error, and then the program will see what missing pieces of information (here value of the t-statistic) can be filled in. It is a good idea, but the screen was packed with possibilities that I fear would overwhelm a beginning student. Alfred Bork, a physicist who has been a pioneer in computer assisted instruction, offers software designers a lot of advice (see [1]) but one of his favorite maxims is to keep the screen simple and relatively uncluttered. This regression screen would not pass Bork's test.

Lastly, it would be helpful if a software designer could permit the instructor to remove unwanted exercises. It can only produce confusion to have a variety of options before the student which are irrelevant to the particular subjects the student is trying to master.... All of the above reservations notwithstanding, I think there is a place for this kind of software and Stat Helper is a start.

I would like to close this section of my paper by remarking that I have been surprised there hasn't been more integration of computer exercises with the major textbooks. I have seen some attempts, but they strike me as excessively conservative in their

departure from traditional methods. It seems to be popular to tie-in with Minitab perhaps because that product has been around for a long time and has been viewed as generally successful.

---

[6]For some examples of the use of Minitab to run statistical simulations see Dambolena [2]. The Minitab graphics, however, leave more to the imagination than we ought to be willing to settle for these days. Although I believe computer simulations (of, for example, applications of the Central Limit Theorem) are one of the most useful computer add-ons to our statistics courses, the quality of graphics from the Minitab output in Dambolena is noticably inferior to the Macintosh graphics in Stat Helper.

---

3. Applications and Coverage.

A good case can be made for spending a significant chunk of a one semester Economics statistics course on regression theory and applications. Students will encounter regression results in most of the empiri al papers on the reading lists of advanced courses. And even in the required intermediate theory courses it isn't unusual to come across a regression result here or there.

The statistics textbooks do a reasonable job at laying out the basic motivation for running regressions, and they provide some assistance in interpreting results. Usually a few special topics such as dummy variables, logarithms or the use of quadratic regressors are included in advanced chapters. In my experience, however, most students never really become comfortable with applications of the regression theory they learn in the first course. This deficiency can be attacked, I think, in a number of ways; I will discuss three. First, excerpts from actual journal articles in which a regression result is central can be distributed in class and discussed in some detail. These details would include theoretical motivation for the specification of the regression equation, interpretation of the estimated coefficients, and hypothesis tests. Some discussion of what it all means is probably a valuable exercise as well. For example, is a student obliged to believe the results if the coefficients are "significant"?

Second, students can be sent off to construct, run and interpret their own regression problem on a personal computer or a mainframe. Although my experiences with this kind of exercise are discussed at some length in [4], I want to describe one activity here.

Toward the end of the course we give our students a regression forecasting exercise using made-up data. Last semester our students were given a one page description of a problem facing a mythical company called Squash Computer. It seems that Squash had the previous year embarked on a special sales program for their popular Zucchini model; Zucchinis were offered at reduced prices on 62 university campuses. Squash was contemplating extension of the program to an additional eight universities, and they required forecasts of ZUCCHINI sales for the new universities. They had data available for each of the 70 universities on 12 variables, such as SIZE of student body, TUITION, average SAT scores, COST of producing a Zucchini at the nearest Squash factory, and dummy variables on whether or not the university had an ENGINEERING program and whether or not the university had FRATERNITIES. Of course Squash also had data on ZUCCHINI sales for the original 62 universities.

This exercise was a valuable learning experience in that students got some hands-on experience using the regression procedures they had up to then only been reading about. They had to decide not only which variables were relevant to the forecast but also what functional form seemed most appropriate

for their regression forecasting equation.[7]

---

[7]As a footnote I would like to say that developing the Zucchini problem wasn't the programming chore such an exercise used to be. I had to decide what variables to include and then what distribution of values would be plausible for my sample of 70 American universities. Perhaps most important I had to decide which variables ought to be correlated and to what degree. I made up the data on my Macintosh using a random sample generator contained in the Data Desk™ software (described below). The point of including this brief explanation is to suggest that the writing of computer- based assignments need no longer be restricted to programmers with lots of time to develop and debug a complicated program. Some of you might want to try it.

My third pedagogical instrument for teaching regression, which I will discuss here at some length, is the use of computer experiments to exhibit a variety of points about regression applications. Again I have developed these exercises using Data Desk™ software on my Macintosh. I can imagine that assignment sheets could be developed and students sent to a computer lab to

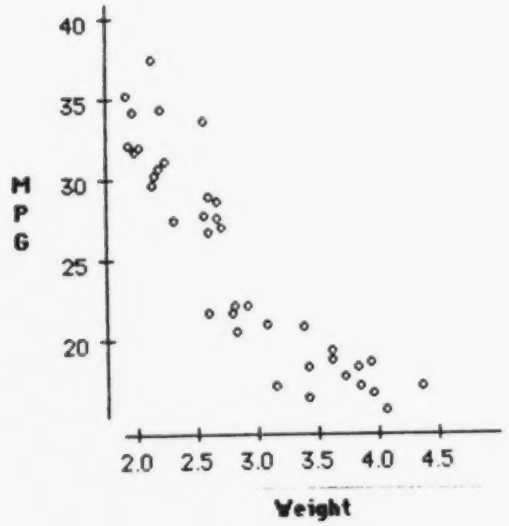
discover a variety of regression "lessons" on their own or that instructors could develop some lecture materials (including handouts or slides—even "hands-on" live demonstrations with a computer and projection equipment in the classroom). These lessons should not be thought of as *mastery* or drill; rather this material would be difficult to include in a textbook. It is instead a use of the computer to expand *coverage* and to do this in a way which makes *applications* more accessible to the average student.

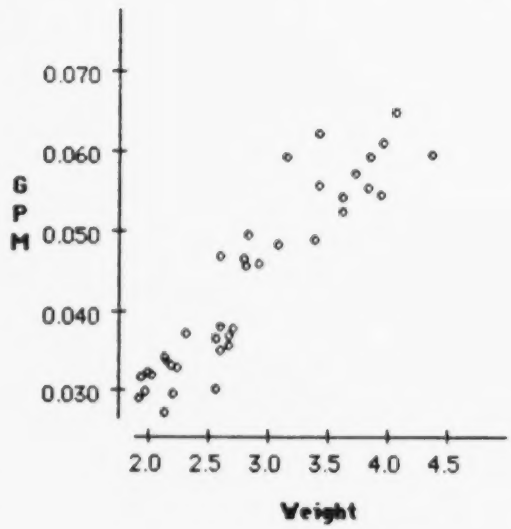
To repeat a line from an earlier paper: "Most students have the naive notion that running a regression, virtually any regression, provides the cutting edge for almost any economic analysis—assuming of course that said regression has 'significant' coefficients and a high R^2." [7, page 83.] We need some materials designed to disabuse students of such notions and to make them more sceptical of regression results. They need to develop some sense about what questions regressions can and cannot be expected to answer. Here is a list of possible examples.

A. Automobile Weight and Fuel Mileage.

Velleman and Velleman [9, see pages 6.9-6.11.] provide some data on automobiles in one of their Data Desk™ data sets

(which they provide for experimentation by students on Macintosh computers). They display the negative relationship between fuel miles per gallon for a car (MPG) and WEIGHT of the car. But a scatter diagram of their data suggests that the relationship is not linear. This isn't surprising; one might think that instead, at least as a first approximation, gallons of fuel consumed per mile (GPM = 1/MPG) ought to be linearly related (proportional) to the automobile's WEIGHT. If you double the weight, you'll double the fuel consumption (per mile). You'll do a better job of capturing the relationship between fuel consumption and weight if you run the GPM = f(WEIGHT) linear regression (see right panel below) instead of the MPG = g(WEIGHT) linear regression (see left panel below). With the GPM regression you'll get a better "fit," i. e. a higher R^2, a higher F statistic, and a scatter of residuals which looks consistent with the assumptions of the regression model (instead of the "U" shaped residuals scatter on the left below). The moral: Think about specifying a functional form that ought to be linear in the relevant variables instead of running the first regression specification that comes to mind.





Dependent variable is: **MPG** **F-ratio** 159
R^2 = 81.6% R^2(adjusted) = 81.0%
s = 2.851 with 38 - 2 = 36 degrees of freedom

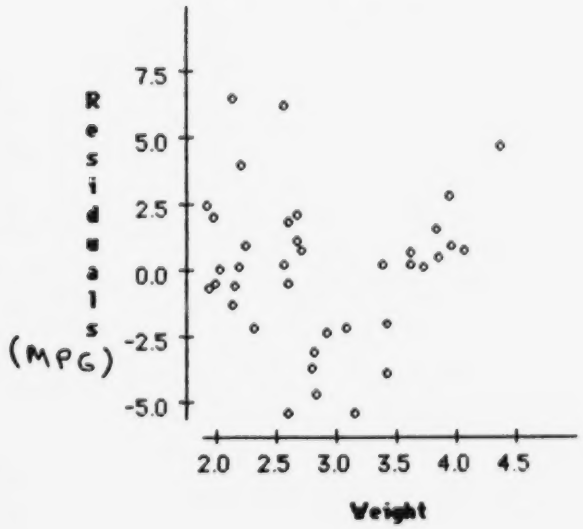
Source	Sum of Squares	df	Mean Square
Regression	1293.52	1	1294
Residual	292.575	36	8.12709

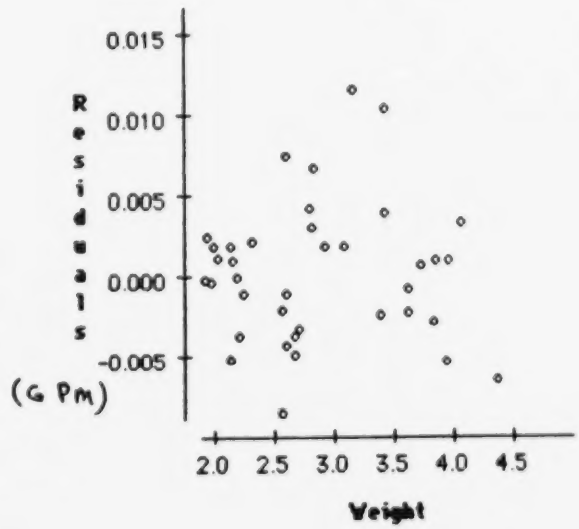
Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	48.7075	1.954	24.9
Weight	-8.36460	0.6630	-12.6

Dependent variable is: **GPM** **F-ratio** 218
R^2 = 85.8% R^2(adjusted) = 85.4%
s = 0.0044 with 38 - 2 = 36 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	0.004242	1	0.004242
Residual	0.000702	36	0.000020

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	-0.000062	0.0030	-0.021
Weight	0.015148	0.0010	14.7

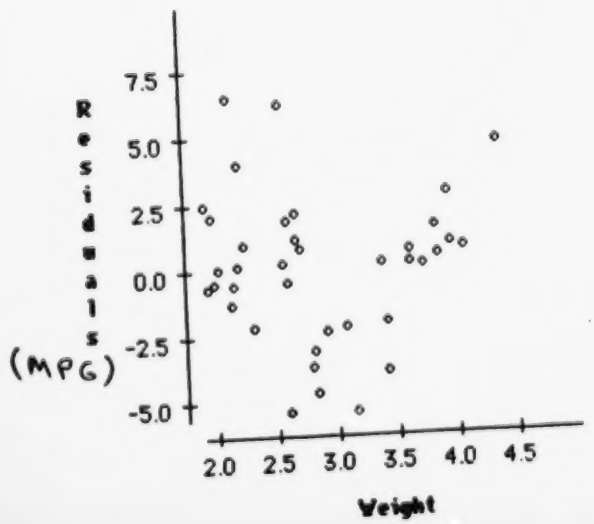




ependent variable is: **MPG**
^{2} = 81.6% R^2(adjusted) = 81.0%
= 2.851 with 38 - 2 = 36 degrees of freedom

ource	Sum of Squares	df	Mean Square	F-ratio
egression	1293.52	1	1294	159
esidual	292.575	36	8.12709	

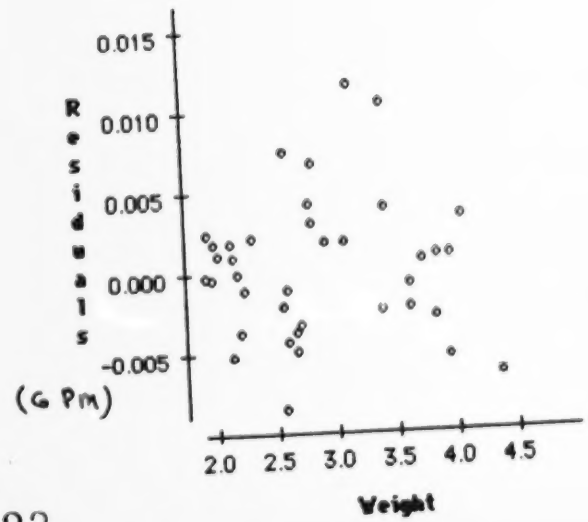
ariable	Coefficient	s.e. of Coeff	t-ratio
onstant	48.7075	1.954	24.9
/eight	-8.36460	0.6630	-12.6



Dependent variable is: **GPM**
R^2 = 85.8% R^2(adjusted) = 85.4%
s = 0.0044 with 38 - 2 = 36 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.004242	1	0.004242	218
Residual	0.000702	36	0.000020	

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	-0.000062	0.0030	-0.021
Weight	0.015148	0.0010	14.7

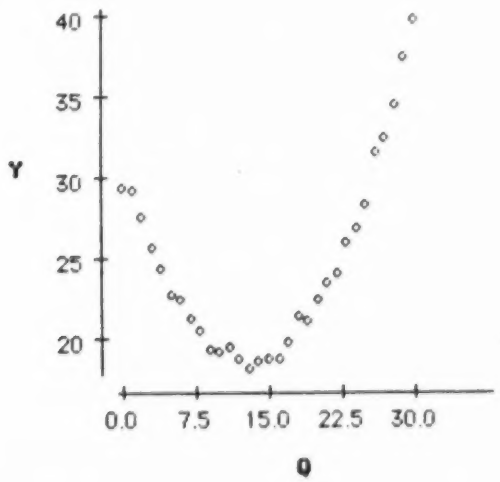


B. Polynomial (quadratic).

For this exercise I generated some data for the following quadratic relationship, perhaps a cost function with average cost Y a quadratic function of quantity Q:

$$Y = 30 - 1.8\,Q + .07\,Q^2 + e$$

where Q varies from 1 to 30 and e is normally distributed with a mean of 0 and a standard deviation of 1/2. The scatter below displays the data pairs I generated; notice it looks like a textbook average cost function. Then I ran two regressions.



On the left below I have (correctly) included both Q and Q^2 as regressors while on the right I have omitted Q^2. The differences are pronounced; the poorer fit and the pattern of the residuals is of course easily seen where Q^2 is omitted (right column of diagrams below). Moreover the correct, non-linear versus the

incorrect, linear predictions (of Y as a function of Q) are also easy to see. Finally the prediction errors are also clear; for the quadratic regression the actual Y's plotted against Predicted (Y's) trace out something resembling a 45^0 line as they should (last diagram in left column) while the regression without the quadratic produces low predictions of Y at both extremes and high predictions in the middle (last diagram in right column).

Dependent variable is: **Y** **F-ratio** 2189
R^2 = 99.4% R^2(adjusted) = 99.3%
s = 0.4834 with 31 - 3 = 28 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	1023.11	2	512
Residual	6.54395	28	0.233712

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	30.3120	0.2445	124
Q	-1.83796	0.0377	-48.7
Qsquared	0.071210	0.0012	58.6

Dependent variable is: **Y** **F-ratio** 7.91
R^2 = 21.4% R^2(adjusted) = 18.7%
s = 5.281 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	220.763	1	221
Residual	808.887	29	27.8927

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	19.9865	1.852	10.8
Q	0.298358	0.1061	2.81

22a.

endent variable is: Y
= 99.4% R^2(adjusted) = 99.3%
0.4834 with 31 - 3 = 28 degrees of freedom

F-ratio
2189

ırce	Sum of Squares	df	Mean Square
ression	1023.11	2	512
idual	6.54395	28	0.233712

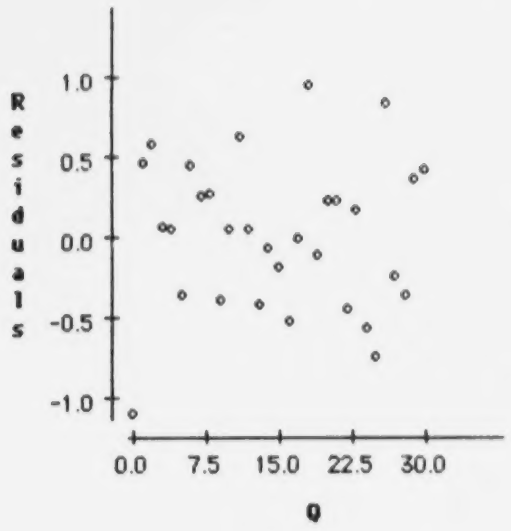
·iable	Coefficient	s.e. of Coeff	t-ratio
stant	30.3120	0.2445	124
	-1.83796	0.0377	-48 7
uared	0.071210	0.0012	58.6

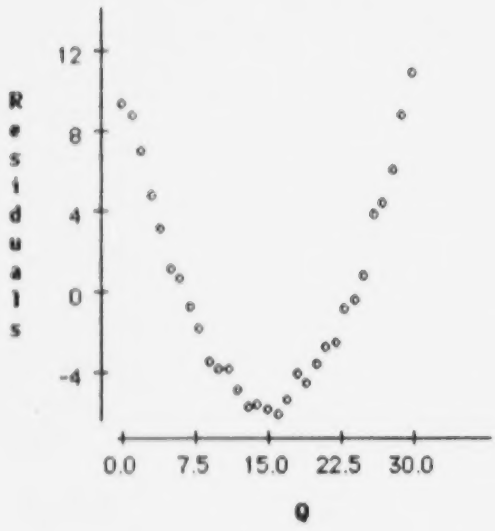
Dependent variable is: Y
R^2 = 21.4% R^2(adjusted) = 18.7%
s = 5.281 with 31 - 2 = 29 degrees of freedom

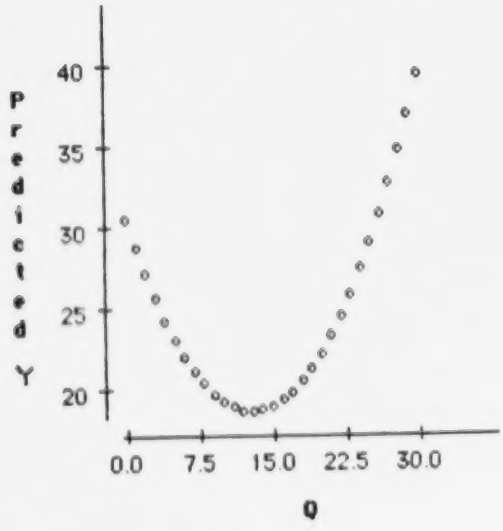
F-ratio
7.91

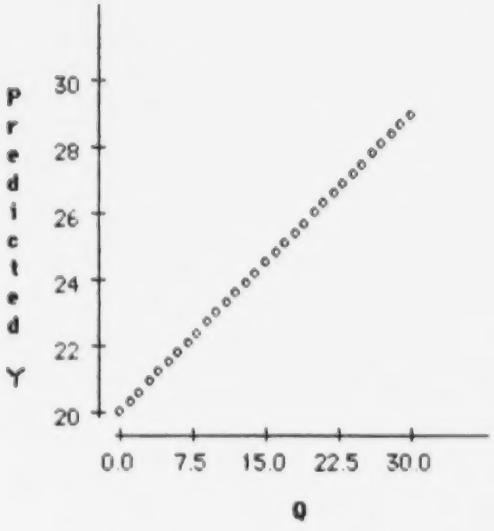
Source	Sum of Squares	df	Mean Square
Regression	220.763	1	221
Residual	808.887	29	27.8927

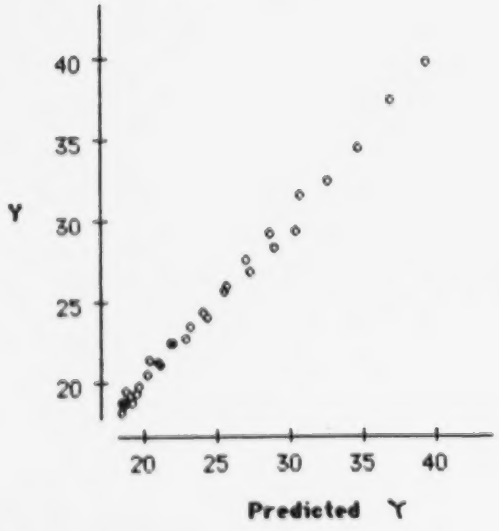
Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	19.9865	1.852	10.8
Q	0 298358	0 1061	2.81

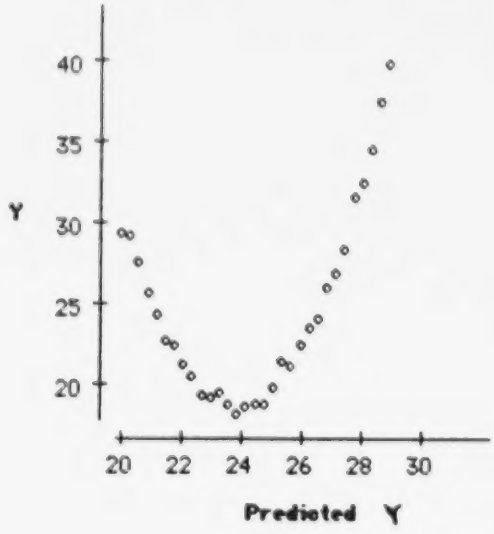

Residuals
1.0
0.5
0.0
-0.5
-1.0
0.0
7.5
15.0
22.5
30.0
Q


Residuals
12
8
4
0
-4
0.0
7.5
15.0
22.5
30.0
Q


Predicted Y
40
35
30
25
20
0.0
7.5
15.0
22.5
30.0
Q


Predicted Y
30
28
26
24
22
20
0.0
7.5
15.0
22.5
30.0
Q


Y
40
35
30
25
20
20
25
30
35
40
Predicted Y


Y
40
35
30
25
20
20
22
24
26
28
30
Predicted Y

C. Omitted Independent Variable.

Next I generated data for the following three-variable relationship:

$$Y = 50 + .8\ X_1 + .5\ X_2 + e$$

where e is normally distributed with a mean of 0 and a standard deviation of 10. The values for the variable X_1 and X_2 were generated at random and independently from a uniform distribution of integers between 0 and 99. Two regressions were run. First, the correctly specified regression (on the left below) and second (on the right) a regression omitting the X_2 regressor. The results are as expected. The fit is poorer when an important variable is omitted. This can be seen both by a lower R^2 and a lower F; also the scatter of residuals shows much larger deviations around zero.

Dependent variable is: Y — F-ratio
R^2 = 88.1% R^2(adjusted) = 87.2% — 103
s = 11.56 with 31 - 3 = 28 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	27564.8	2	13782
Residual	3739.54	28	133.555

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	42.6582	6.802	6.27
X1	0.889653	0.0697	12.8
X2a	0.541060	0.0817	6.62

Dependent variable is: Y — F-ratio
R^2 = 69.3% R^2(adjusted) = 68.3% — 65.6
s = 18.19 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	21707.0	1	21707
Residual	9597.27	29	330.940

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	78.6550	6.438	12.2
X1	0.888813	0.1097	8.10

24a

endent variable is: Y

= 88.1% R^2(adjusted) = 87.2%

11.56 with 31 - 3 = 28 degrees of freedom

rce	Sum of Squares	df	Mean Square	F-ratio
ression	27564.8	2	13782	103
dual	3739.54	28	133.555	

iable	Coefficient	s.e. of Coeff	t-ratio
stant	42.6582	6.802	6.27
	0.889653	0.0697	12.8
	0.541060	0.0817	6.62

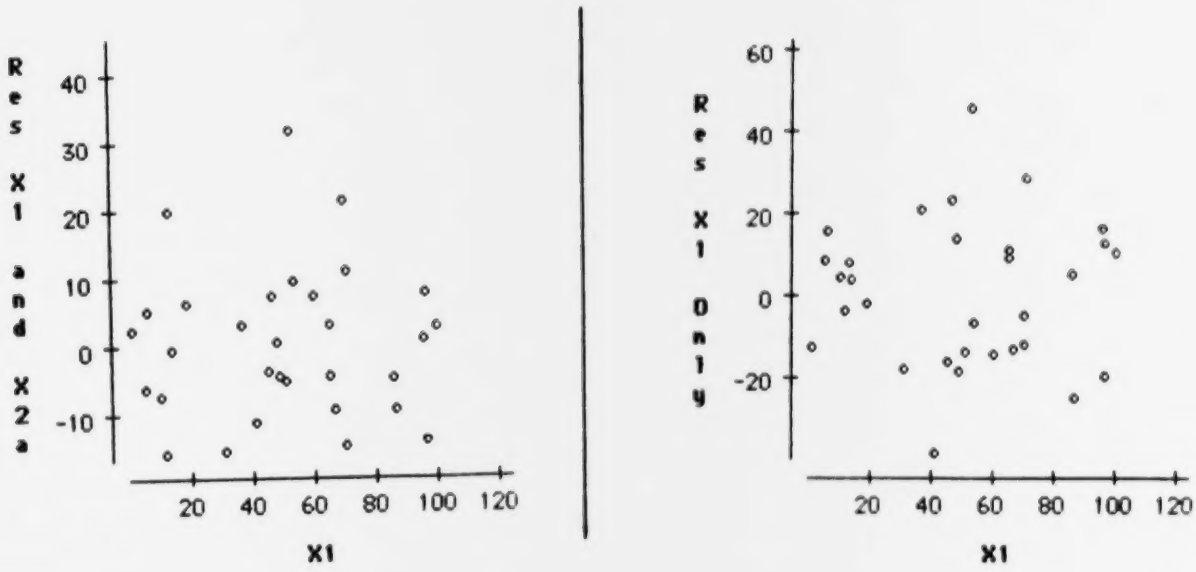
Dependent variable is: Y

R^2 = 69.3% R^2(adjusted) = 68.3%

s = 18.19 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	21707.0	1	21707	65.6
Residual	9597.27	29	330.940	

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	78.6550	6.438	12.2
X1	0.888813	0.1097	8.10



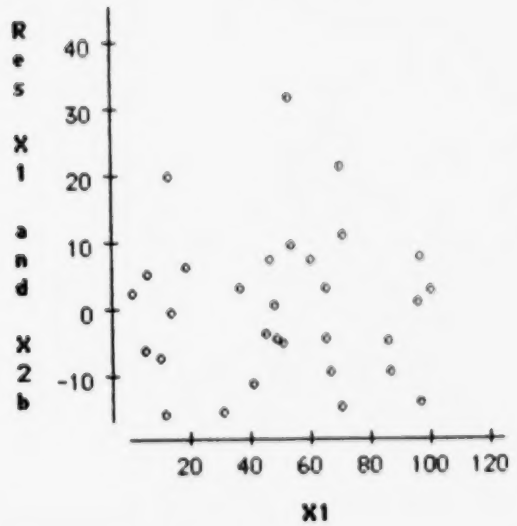
Then I repeated this exercise with one important change. A new X_2 variable was obtained by taking, for each of the 31 observations, the mean of X_1 and X_2 from the first regression. (To distinguish my output the uncorrelated case above shows the second variable as X2a while the correlated case below uses X2b.) This change produces high, positive correlation between (the original, unchanged) X_1 and the new X_{2b}. The regression results with (on the left) and without (on the right) the regressor X_{2b} are shown below. Now the cost of omitting X_{2b}, in terms of lower R^2 and higher variability of the residuals, is much reduced. The effect of multicollinearity on t-ratios is evident: the t-ratio falls

sharply for the X_{2b} coefficient for the regression on the left. For the regression on the right the estimate for the X_1 coefficient increases substantially above its true value; this reflects the positive bias caused by omitting the positively correlated X_{2b} regressor.

Dependent variable is: **Y** **F-ratio**
R^2 = 90.9% R^2(adjusted) = 90.3% 140
s = 11.56 with 31 - 3 = 28 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	37355.1	2	18678
Residual	3739.54	28	133.555

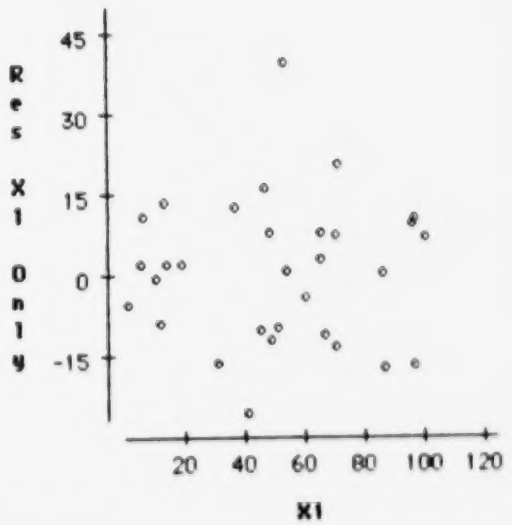
Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	42.6582	6.802	6.27
X1	0.848593	0.1073	7.91
X2b	0.582120	0.1634	3.56



Dependent variable is: **Y** **F-ratio**
R^2 = 86.8% R^2(adjusted) = 86.3% 190
s = 13.69 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	35659.9	1	35660
Residual	5434.67	29	187.402

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	62.0225	4.845	12.8
X1	1.13920	0.0826	13.8

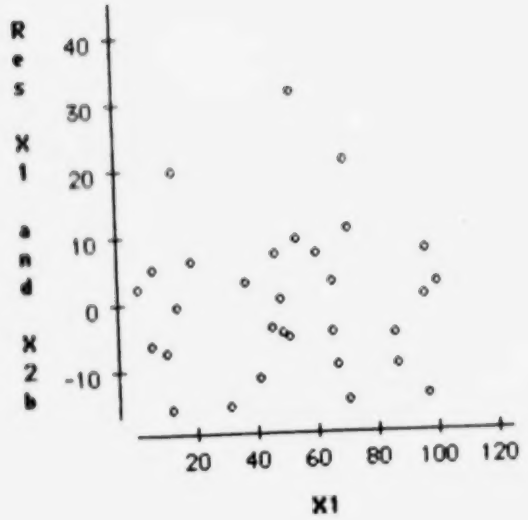


26a

dent variable is: Y
90.9% R^2(adjusted) = 90.3%
1.56 with 31 - 3 = 28 degrees of freedom

ce	Sum of Squares	df	Mean Square	F-ratio
ssion	37355.1	2	18678	140
ual	3739.54	28	133.555	

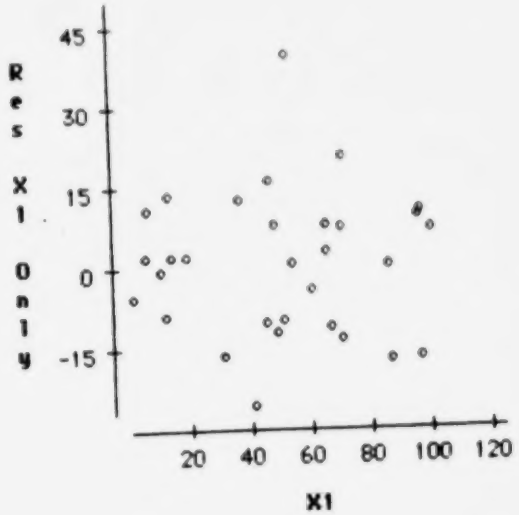
able	Coefficient	s.e. of Coeff	t-ratio
ant	42.6582	6.802	6.27
	0.848593	0.1073	7.91
	0.582120	0.1634	3.56



Dependent variable: Y
R^2 = 86.8% R^2(adjusted) = 86.3%
s = 13.69 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	35659.9	1	35660	190
Residual	5434.67	29	187.402	

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	62.0225	4.845	12.8
X1	1.13920	0.0826	13.8



D. Logarithmic Relationship.

For my last example I consider a demand relationship between PRICE and QUANTITY, namely:

$$QUANTITY = 3000\ (PRICE)^{-.5}\ (exp)^{u}$$

where u is a multiplicative, normally distributed error term with a mean of zero and a standard deviation of .05. PRICE varied from 10 to 40 in steps of 1. The scatter of QUANTITY and PRICE is displayed below on the left and the scatter of ln(QUANTITY) and ln(PRICE) on the right. Note that the scatter between QUANTITY and PRICE is non-linear as expected; also the residuals plotted against PRICE suggest nonlinearity. The fits as reflected by the R^2s also shows that the logarithmic version works better. Finally when the ln regression residuals are transformed to QUANTITY units (by calculating QUANTITY - exp(Predicted lnQUANTITY)) and plotted against PRICE the nonlinearity in PRICE residuals disappears (last diagram, center).

Dependent variable is: **QUANTITY** **F-ratio**
R^2 = 85.8% R^2(adjusted) = 85.3% 175
s = 52.21 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	477160	1	5e+5
Residual	79041.7	29	2725.58

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	976.299	27.84	35.1
PRICE	-13.8709	1.048	-13.2

Dependent variable is: **lnQUANTITY** **F-ratio**
R^2 = 97.0% R^2(adjusted) = 96.9% 923
s = 0.0358 with 31 - 2 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square
Regression	1.18299	1	1.1830
Residual	0.037153	29	0.001281

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	7.97072	0.0513	155
lnPRICE	-0.491556	0.0162	-30.4

2-72

endent variable is: **QUANTITY** **F-ratio**
= 85.8% R^2(adjusted) = 85.3% 175
52.21 with 31 - 2 = 29 degrees of freedom

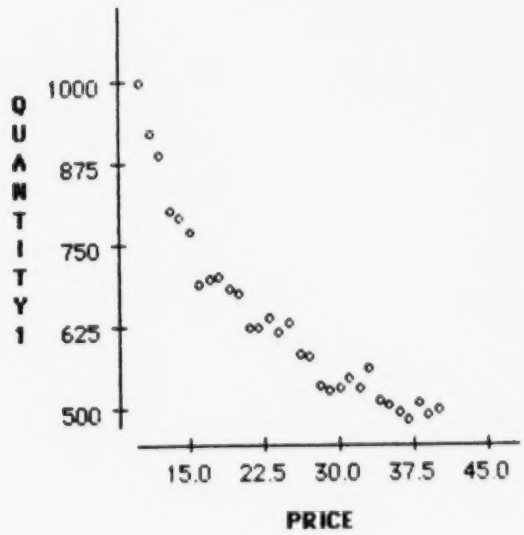
irce	Sum of Squares	df	Mean Squ:re
ression	477160	1	5e+5
idual	79041.7	29	2725.58

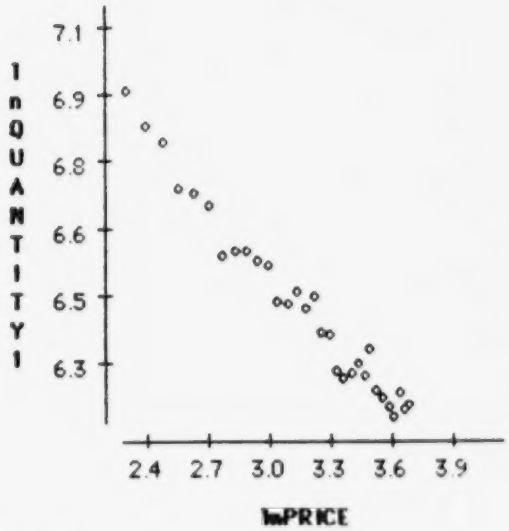
iable	Coefficient	s.e. of Coeff	t-ratio
stant	976.299	27.84	35.1
CE	-13.8709	1.048	-13.2

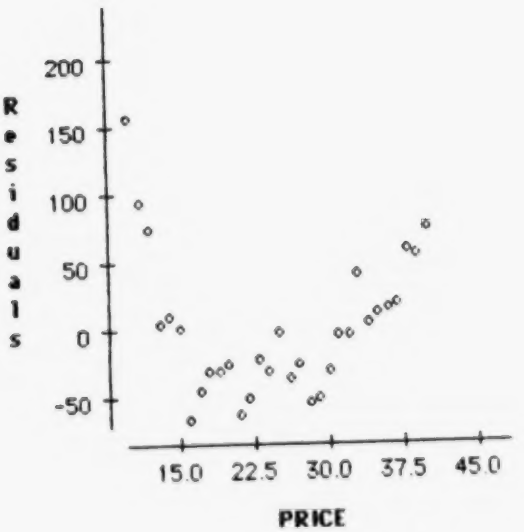
Dependent variable is: **lnQUANTITY** **F-ratio**
R^2 = 97.0% R^2(adjusted) = 96.9% 923
s = 0.0358 with 31 - 2 = 29 degrees of freedom

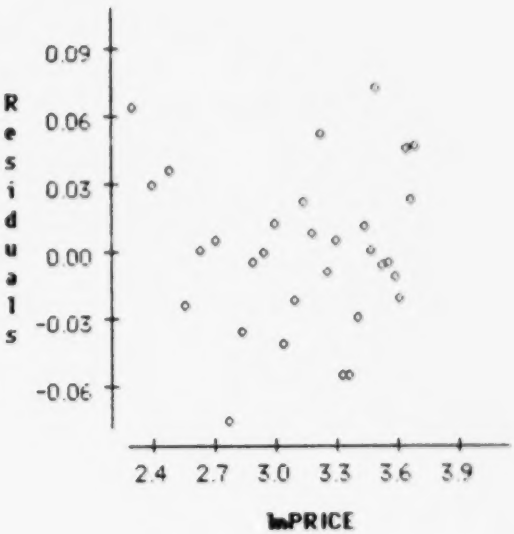
Source	Sum of Squares	df	Mean Square
Regression	1.18299	1	1.1830
Residual	0.037153	29	0.001281

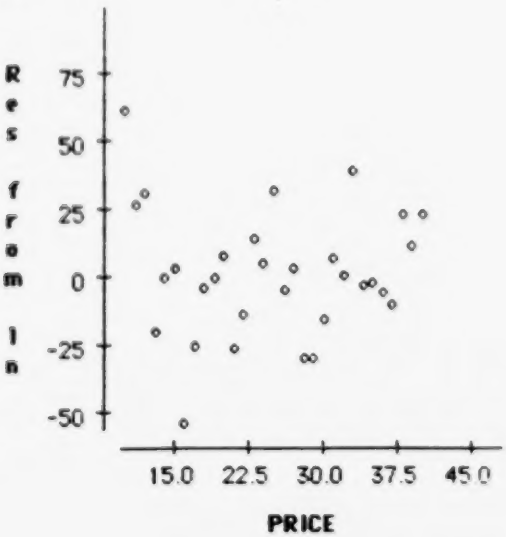
Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	7.97072	0.0513	155
lnPRICE	-0.491556	0.0162	-30.4


QUANTITY1
1000
875
750
625
500
15.0
22.5
30.0
37.5
45.0
PRICE


lnQUANTITY1
7.1
6.9
6.8
6.6
6.5
6.3
2.4
2.7
3.0
3.3
3.6
3.9
lnPRICE


Residuals
200
150
100
50
0
-50
15.0
22.5
30.0
37.5
45.0
PRICE


Residuals
0.09
0.06
0.03
0.00
-0.03
-0.06
2.4
2.7
3.0
3.3
3.6
3.9
lnPRICE


Resfrmln
75
50
25
0
-25
-50
15.0
22.5
30.0
37.5
45.0
PRICE

4. Conclusion.

The reader might be thinking that all of the above exercises are relatively simple and straightforward. That's precisely the point I want to make. I recall that before the current generation of software was available it took me perhaps a month or two to put together a FORTRAN program to do all the above (and admittedly a bit more) for the mainframe. And then PC's became available, and I had to invest a few more days because the mainframe FORTRAN wasn't compatible with the FORTRAN for the PC's. After I had converted the FORTRAN code, I discovered I had been using a library program for the mainframe that had no counterpart on my PC. Most of us will think twice before making a comitment to develop such programs on our own. But now, I want to argue, you can spend a day or two—perhaps less if you have a clear view of exactly what you want to do—and at the end of that period you'll have some good lecture materials for your statistics class. Further, after you have figured out how the software works, subsequent exercises will take even less time. The purpose, I should remind you, is to advance our students' understanding of some of the pitfalls to be avoided in constructing regression models. They should develop

some appreciation of possible reasons why regression results published in professional journals are sometimes confusing or inconclusive. All of this would I think be difficult to glean from the pages of textbooks.

I would like to end with two caveats. Software doesn't always perform as expected or even as advertised. Invoking an operation that seems to make sense and that isn't defined to be illegal in the manual doesn't necessarily accomplish what reasonable people might guess it will accomplish. And there may be no warning flags to alert you to possible mistakes.

Finally, beware of so-called random number generators; I have been told that personal computer generators often cannot pass tests for randomness. In constructing the last example above I had considerable trouble producing the results theory led me to expect. I repeated the exercise several times—thinking I must have made some errors in linking the operations. Eventually I looked at my normal error distribution (see below). Notice that the two normal error distributions (generated with $\mu = 0$ and $\sigma = .05$) each have three observations more than two standard deviations below the mean, and they all occur with values of

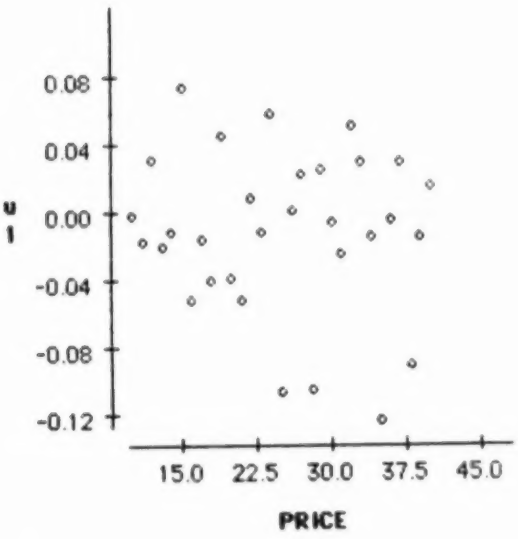
PRICE which are well above average.[8] With these random shocks

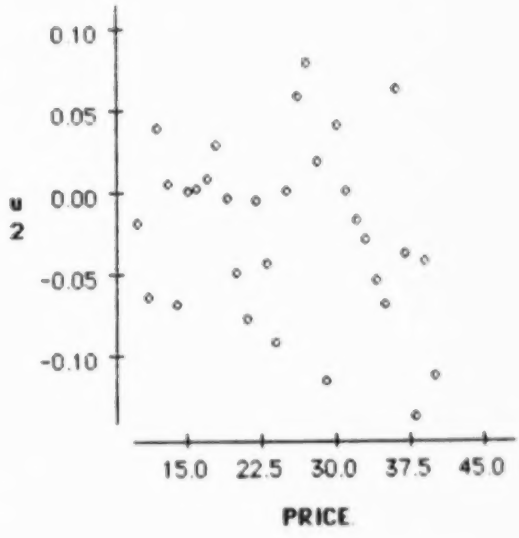
---

[8]I haven't performed any tests to check whether these error distributions would fail a normality test. Either way a bad draw on random errors will often produce results which don't demonstrate the theoretical point you want to make.

---

you won't be surprised to learn that my slope coefficients were way off target on the low side.





Of course now I can take some of these puzzling regression results and show them to my class. And now I have an explanation. Every cloud has its silver lining.

References

[1] Alfred Bork, Learning with Personal Computers, Harper & Row, 1987.

[2] Ismael G. Damolena, "Using Simulation in Statistics Courses," Collegiate Microcomputer, Vol. 4, no. 4 (November 1986), 339-344.

[3] D. P. Doane, Exploring Statistics with the IBM PC. Reading, MA: Addison-Wesley, 1985.

[4] F. Trenery Dolbear, Jr., "The Use of Microcomputers in Economics Statistics Courses," Social Science Microcomputer Review, 4:4, Winter 1986, 438-456.

[5] S. M. Dowdy, Statistical Experiments Using Basic. Boston: Duxbury, 1986.

[6] Dennis Dunn and Mary Ann Atkinson, Stat Helper I and II, Santa Barbara, CA: Kinko's Academic Courseware Exchange.

[7] L. B. Pulley and F. T. Dolbear, "Computer Simulation Exercises for Economics Statistics," The Journal of Economic Education, 15 [Winter 1984], 77-87.

[8] Marc M. Sebrechts and Lael J. Schooler, "An Artificial Intelligence Approach to Analyzing Student Errors in Statistics," Collegiate Microcomputer, Volume 1 no 1 (February 1987), 82-86.

[9] Paul F. Velleman and Agelia Y. Velleman, The Data Desk Handbook, Data Description, Inc., Box 4555, Ithaca, NY 14852.